Learning Spatial Representation for Safe Human-Robot Collaboration in Joint Manual Tasks

Mohammad Ali Zamani[1], Hadi Beik Mohammadi[1], Matthias Kerzel[1], Sven Magg[1] and Stefan Wermter[1]

***Abstract*— Programming robots for a safe interaction with humans is extremely complex especially in collaborative tasks. One reason is the unpredictable behaviour of humans that may have an intention which is not clear to the robot. We present a novel architecture for a safe human-robot collaboration scenario in a shared tabletop workspace based on intuitive multimodal language and gesture instructions and behaviour recognition. In our example scenario, a human and a robot arm collaboratively have to assemble a Tangram puzzle. The configuration space of the robot is constrained by a combination of learned behaviour patterns of the user by tracking its arm and direct audio-visual instructions regarding the sharing of the workspace. This ensures a safe and non-obstructive collaboration behavior of the robot which can constantly be updated during task execution. In this paper, we present initial results with a focus on instruction understanding.**

I. INTRODUCTION

In the future, robots are expected to work as companions to humans in various areas ranging from domestic to care-giving scenarios. Ensuring human-robot interaction safety is one of the major challenges for realizing this goal [1]. How this safety can be achieved is still a research challenge [2]. Even with well-engineered robots, it would be unrealistic to move robots directly from factories to home environments to perform complex tasks such as care-giving [3], [4]. Moreover, the robots also have to adapt to new environments to avoid hazardous actions since using experts to program a robot for every possible environment is not feasible. Hence, the need for adaptive learning algorithms arises.

Spoken language can be considered as one of the most effective communication channels to warn robots about threats. For example, robots may not notice an external threat or misplanning that may harm the robot or even a human. However, a human can easily warn or guide the robot by a proper verbal utterance toward a safer interaction.

It is difficult for robots to learn to avoid a threat by only verbal warnings from a human user. Especially if the warning includes spatial information, deictic gestures are more intuitive. For example, when a human and a robot share an operational space or the human wants to limit the workspace of the robot for safety reasons, combining spoken instruction with a pointing or drawing gesture becomes more efficient. This leads to a safer and disturbance-free sharing of mutually utilized space.

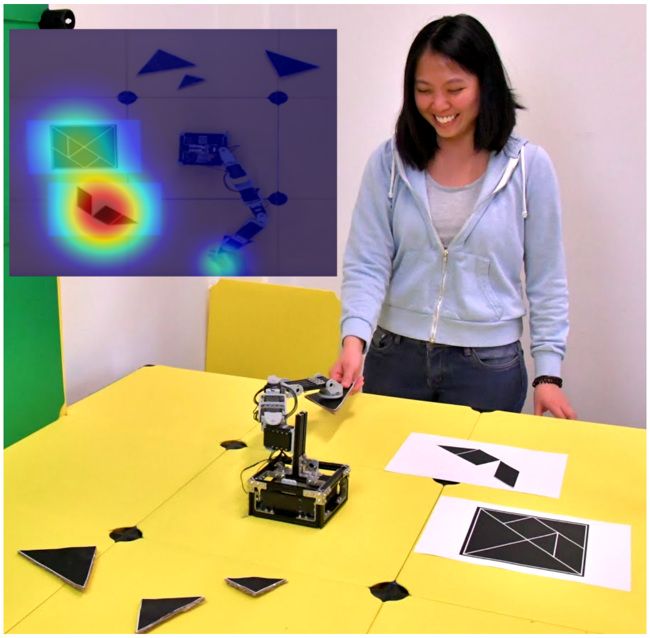


Fig. 1: A person is solving a tangram puzzle in collaboration with a robot arm. The robot arm is instructed to avoid the person's workspace while fetching puzzle pieces from the far end of the table. The top left image shows the top view overlaid with spatial representation which can be learned by interaction with the user. The robot plans its motion incorporating the adaptive spatial constraints. The area types range from fully-restricted (red) to non-restricted ones (blue).

To realize this, we propose an architecture for the intuitive instruction of a robotic collaborator with speech and gestures with a special focus on safety. We demonstrate the proposed architecture in a human-robot interaction scenario, where a robot and a human jointly solve the collaborative task of building a Tangram puzzle[1]. Both the robot and human operate with their hand and end-effector in a shared, table-sized environment. While both use this space to achieve a joint goal, it is important that the robot does not cause harm to the human through collisions or impede the human's actions by blocking access to parts of the workspace.

The robot motion planning is based on a spatial representation (SR) of the workspace. For a safer collaboration, a robot assistant can learn to refine the SR by interaction with the user. The SR can be defined as a mapping from

*The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169) and the European Union under project SECURE (No 642667).

[1] Knowledge Technology, Department of Informatics, University of Hamburg, Germany. zamani, 6beik, kerzel, magg, wermter @informatik.uni-hamburg.de

[1] Tangram is popular dissection puzzle where seven basic shapes are used to construct more elaborate shapes.

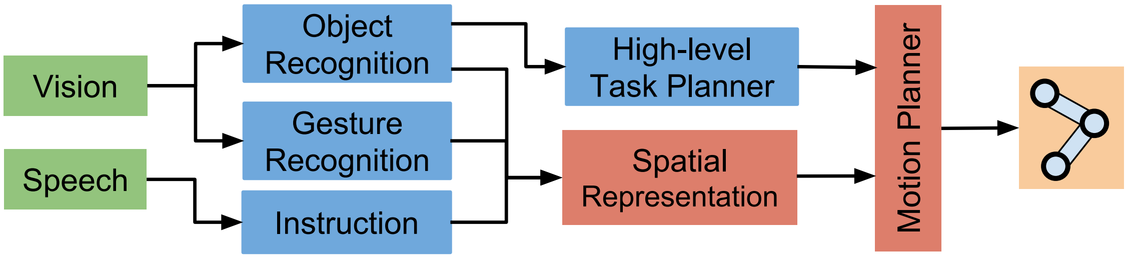


Fig. 2: Architecture for safe human-robot collaboration through speech and gesture instructions.

the 2D/3D workspace to a value that represents a constraint for the motion planning. As shown in figure 1 (the top-left corner), it can range from fully restricted areas (red) over partially restricted areas (yellow) to a non-restricted (blue) areas. In the case, that the robot needs to violate the user's preferences to carry out an action, the robot should pass over less restricted areas (yellow). This can later be extended to different interaction areas, e.g. do not enter vs do not place objects.

The spatial representation can be extracted from multi-modal instructions such as gesture and language; it can also be understood from experiences the robot makes during an ongoing interaction. A combination of instructed and learned behaviors will create the most intuitive collaboration between robot and human. In this paper we focus on the first part of this collaborative interaction, the explicit instruction of the robot: Restricted areas can be marked explicitly by gestures and spoken instructions (e.g., "*do not enter this area*"). The human user draws regions in the workspace using deictic gestures and simultaneously describes the functionality of the regions. The instructed functionality is interpreted as whether the robot can enter the region. These constraints are then, in turn, considered during the robot's path planning.

Safety issues can also occur during a given task. An example is an unexpected external (environmental) event or the extension of the human's workspace. This case is difficult because it requires continuous updates to the robots internal representation of the world. Moreover, a warning utterance which has all the necessary information to take an action is not always expected to be received from the human. These warnings are also mostly given in stressful conditions where humans may not follow predefined messaging conventions (e.g. it may lack proper syntax of written language, e.g. "*Be careful!*"). The robots can modulate their actions (e.g. slow down, use other objects, or temporarily stop) if they can understand the intention of the given warning.

In both explicit and implicit instructions, situations can arise in which not all constraints can be satisfied. In such a case constraints are relaxed based on their type and priority to ensure that the robot partner can still actively collaborate. This violation of instructions, however, is reflected in the robots behavior. If the robot, for instance, reaches into the workspace of the human, it will decrease its velocity or announce its action via a dialogue system.

We present initial work on a full realization of a safe human robot collaboration during a joint manual task. We present details on the system architecture, initial results from gesture and speech recognition as well as motion planning, and a physical experimental setup.

II. RELATED WORK

A. Language Instructions for Safety

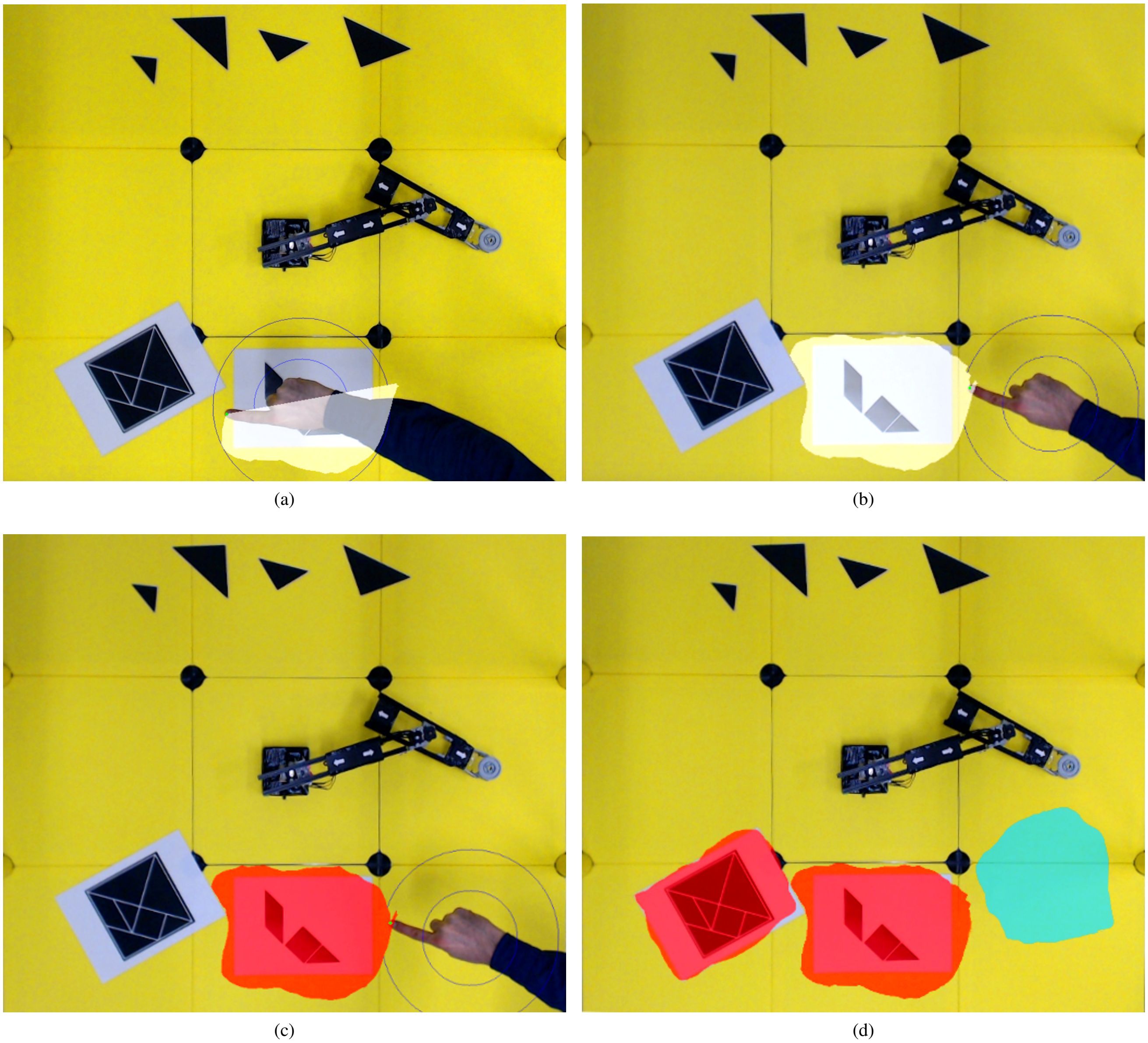
How robots react to safety warnings is not addressed exhaustively in the literature. The closest related research area is assigning tasks to robots by verbal instructions. Lauria et al. [5], [6] and Nishizawa et al. [7] apply rule-based methods to utilize spoken language instructions which can cover only a limited number of scenarios.

The robots can modulate their actions (e.g., slow down, use other objects, or temporarily stop) if they can understand the intention of the given warning. This is discussed to some extent by Hua et al. [8] with a focus on applications like web search. Another project in Facebook DeepText [9] was started to connect users to their products inspired by Collobert et al. [10]. Abdulkader et al. [9] capture the user's intention in a unique phrase. For example, sentences like "*I need a ride*", "*Take a cab*" or "*But, I need to take a taxi*" are interpreted as "Request a ride".

The detection of human intention is critical to safety issues which can lead to a proper reaction of the robot and ideally helps it to update the robots internal representation of the world. Zamani et al. [11] implemented a hybrid system which detects the intention of the given instruction in a predefined set of tasks in a domestic scenario (e.g. "boil water" or "clean the living room"). Then, a deep reinforcement learning planned the sequence of steps to fulfill the task depending on the environmental state. Recognizing the acoustic emotion can also provide extra clues about safety issues. Lakomkin et al. [12] developed an approach called EmoRL that rapidly recognizes the angry state of the human speaker. EmoRL is trained by optimizing the latency and accuracy concurrently enables a continuous emotion recognition which can be used for emergency stops.

B. Models for Socially Adequate Robot Behavior

For a robot to behave truly socially acceptable and safe, its behavior should not be based on following situational



(a)

(b)

(c)

(d)

Fig. 3: Areas designated by gesture and speech instructions. While "drawing" areas on the table the user instructs the robot about expected behavior: (a) to (c): *"Here I will build the puzzle, please do not enter this area."* (d) *"And here I put the pattern that I want to build, please do not occlude it."* and *"Hand me the puzzle pieces in this area."*

commands, but on a deeper understanding of the joint collaborative task and the needs of the human user. For instance, a robot should be aware of the human-preferred workspace and should follow the general rule to not obstruct this area without the need to be "reminded" of this by repeated user instructions. Moreover, deep reinforcement learning approaches have proved they are capable of end to end learning to control a robot regardless of mechanical structure [13] and with high number of constraints and singularities [14] in a collaborative dynamic environment [15]. Several formal models for socially adequate robot behavior have been proposed. For instance, Lindner [16] suggested a conceptual model for social spaces for mobile robots.

III. METHODOLOGY

A. Architecture

Our proposed architecture is shown in Figure 2. The robot learns to constrain its motion planning based on the human instructions given through speech and gestures. The overall architecture is not task-dependent, and the motion planner can be applied for various robot models and tasks. The robot performs a given task based on the spatial representation which is used to constrain the motion planning. The SR is updated with new instructions and therefore, the robot replans its motion. The user can use gestures to define areas in the workspace and simultaneously spoken instructions are used to label these areas according to the instruction.

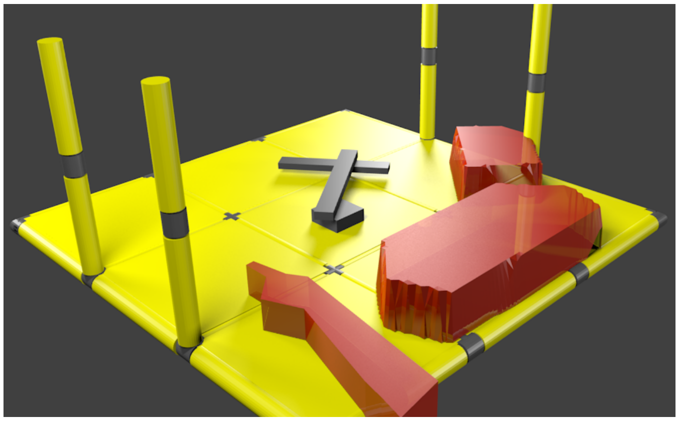


Fig. 4: Visualization of the configuration space of the robot, constrained areas that are blocked for motion planning are realized as physical obstacles (shown in red).

The architecture introduces an adaptive spatial representation where new areas can be added and removed during operation.

The high-level planner will create a series of goals for the robot's end effector. For example, which piece of puzzle should be picked next. Anyway, we do not go into detail about the task-specific high-level planning. We rather assume that the robot needs to move its end-effector to a series of locations in the workspace. Instead, we focus on the question: How can the robot realize these motions without or with minimal violation of the constraints given in the instructions?

B. Gesture Recognition

In order to give the robot instructions about certain parts of the shared workspace, the human uses gestures to "draw" these areas while giving verbal instructions. The vision module observes the shared workspace using input from a ceiling mounted camera. The user's hand is tracked by color selection and applying several filters in the video stream. The module detects the tip of the index finger by finding the farthest point from the centroid of the hand and detects if the index finger is extended for a drawing gesture. Figure 3 (a-c) shows the detected hand and index finger. If the index finger is extended far enough, the module recognized a drawing gesture and tracked the tip of the index finger to identify designated areas in the workspace. Meanwhile, the speech module and the connected module for instruction understanding assign a behavior to the drawn area.

C. Spoken Instructions

The user instructions together with gesture recognition form the spatial representation for the motion planning. The semantic representation of instructions is mapped into a value from 0 to 1 which is ranging from non-restrictive to fully restrictive instructions. This value shows the constraint level for the area which is shown as a radial basis function with the same center of the marked area by gesture and the maximum value obtained from the instruction (see Figure 5).

Keyword spotting or Bag of Words (BOW) [17] can be used to evaluate the instruction. However, those methods are limited when the instruction is complex, also requiring feature extractions and engineering to work on a larger scale. To overcome the complexity of recognizing spoken instructions, a specific corpus can be collected and then trained by state-of-the-art document classification models [18].

A corpus with continuous annotation of 0 to 1 or ordinal classes (from non-restrictive to fully restrictive) can be used for training the instruction module. The instruction can be both explicit ("*You are not allowed to enter this area*") and implicit ("*Only I can work here*") instructions. When there is no restrictive instructions, the robot can enter the area. The permission is either clearly granted in the instruction or the given instruction does not forbid the reachability.

The robot may receive instructions during the operation as well. For example, a short instruction such as "*Not there!*" without any gesture means it refers to the current position of the robot in the workspace. Therefore, this point can be added to the spatial representation.

D. Motion Planning with Collaboration Constraints

A high-level action planner decides what actions the robot should carry out to perform a given task. For a collaborative manual task, the robot can adopt the role of a helper by fetching items for the human user; this high-level action planner is expected to create a sequence of pick and place actions. The execution of individual pick and place actions is controlled by a motion planner that is constrained by the previously given instructions.

The motion planner is realized by first importing a URDF model of the robot and its static environment (e.g., the table top) into a simulation environment. Movable, dynamic objects (e.g., the Tangram pieces) are then added to this simulated environment. Their position and orientation are supplied by the vision module. Finally, the full planning landscape is created by realizing spatial representation as physical obstacles. Figure 4 shows the simulated environment that is used for planning. Three red obstacles indicate the areas which the robot arm should avoid: the working area of the human at the close end of the table, the human's arm, which is currently on the left side of the human's working area and the area in which the human placed a document that is needed during the task (e.g., a recipe or manual). These obstructing areas are created as 3d-height maps from pixel data from the vision module and added to the simulation environment. They serve, like any other physical object, as an obstacle for trajectory planning. The Robot Operating System (ROS) ROS trajectory planner MoveIt! is used in conjunction with the inverse kinematics solver TRAC-IK [19] to compute suitable motion trajectories for each pick or place action. Due to the obstacles in the simulation environment that represent the given instructions as well as the human arm, collision with the human as well as obstruction of the assigned working areas is prevented.

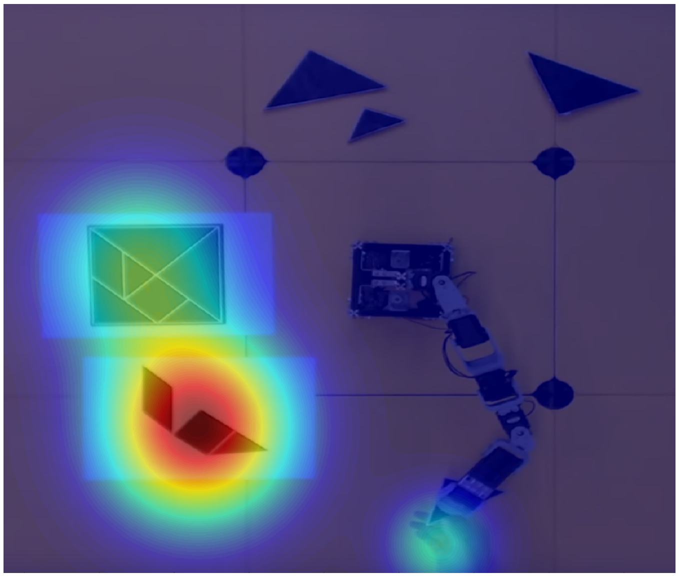


Fig. 5: The projected heat map on the robot's workspace. The heat map values are between 0 (blue) and 1 (red) which range from the non-restricted to the fully restricted areas.

IV. IMPLEMENTATION AND RESULTS

In this section we present the preliminary implementation of the proposed architecture for safe human-robot collaboration in a joint manual task scenario, which was used for initial trials. A demonstration video for the system can be found at: https://youtu.be/-6gmSfjL02I.

A. Technical Realization

The robot arm in our experimental setup has four degrees of freedom and is mounted centrally in a table. A scaffolding structure on the sides is used to mount a ceiling camera that provides a top view of the tabletop. The robot arm is articulated by Dynamixel servomotors that are controlled via the PyPot framework [2]. The Tangram puzzle pieces are magnetic. The robot arm uses an electromagnetic gripper to pick and place puzzle pieces. The architecture is implemented using Python and OpenCV[3] for the vision and gesture recognition system.

B. Experimental Scenario

In this scenario, the human is going to complete a Tangram puzzle in a shared, table-sized work environment depicted in Figure 3. The pieces of the puzzle are located on the far side of the table that is not reachable for the human. The task of the robot is to bring all parts of the puzzle into the workspace of the human. The setup is shown in Figure 1.

Initially, the human instructs the robot via speech and gestures to avoid the working area of the human in which the puzzle is constructed and also not to occlude the printout that depicts the goal state of the puzzle, as shown in Figure 3 by two different red areas. *"Here I will build the puzzle, please do not enter this area. And, here I put the pattern that I want to build, please do not occlude it."* The robot is programmed to bring the puzzle pieces to the human. Since the location of the handover between robot and human should be ergonomic and comfortable for the user, a handover area can be defined by the user (blue area): *"Hand me the puzzle pieces in this area."*

During operation, the robot will pick up a random puzzle piece from the far side of the table and bring it into the handover area. The motion planning for each trajectory considering the instructions are given by the user, i.e., the designated (red) areas are avoided.

C. Interpretation of the Spatial Representation

We represent a wide range of possible cases from highly constricted to non-constrained by assigning an instruction label to each area designated by a gesture. These labels are described below:

1) Restricted areas: Restricted areas are parts of the workspace that the robot should avoid. Examples might be blocking a workspace The robot should avoid entering these areas in general and of course dropping objects there.

2) Semi-restricted Areas: There are no obstacles in these areas to consider collision avoidance. It is desired that the robot avoids these areas. If necessary, the robot can enter these areas, however, it should not drop any object there. An explicit instruction like "*Do not block this area, I need to see the pattern.*" can create this constraint in the SR. Then, if the planner is restricted to use this area, it should pass quickly to not occlude the pattern too much. Alternatively, if the robot detected the human's hand passing an area often in the past, then the robot should only slowly enter these areas if it is not actively used by the human at the moment, i.e. when the humans hands are not in the areas. Similarly, when a hand-over from the robot to the human is planned. So, it reduces potential risk to harm the human.

3) Non-restricted Areas: The robot can pass and drop item in these areas.

4) Assigned Areas: Finally, assigned areas are the opposite of restrictions: these are areas that are assigned to the robot to carry out a task. If possible, the robot will adhere to these areas for all its actions.

D. Intensity of the Instruction

Not all instructions have the same importance; this is often signaled by the user by using different formulations that carry a different sentiment. We represent this difference in wording by strong sentiment: *"Absolutely do not touch my new smartphone here."* neutral sentiment: *"Here is my cutting board, be careful when I am working here."* weak sentiment: *"If possible try to stay away from the pattern."* The spatial representation can build by the sentiment values and instructions with weaker sentiment will be relaxed first, if a task can not be carried out.

[2] Python library for controlling Dynamixel motors: http://poppy-project.github.io/pypot/

[3] Open Source Computer Vision Library: https://opencv.org/

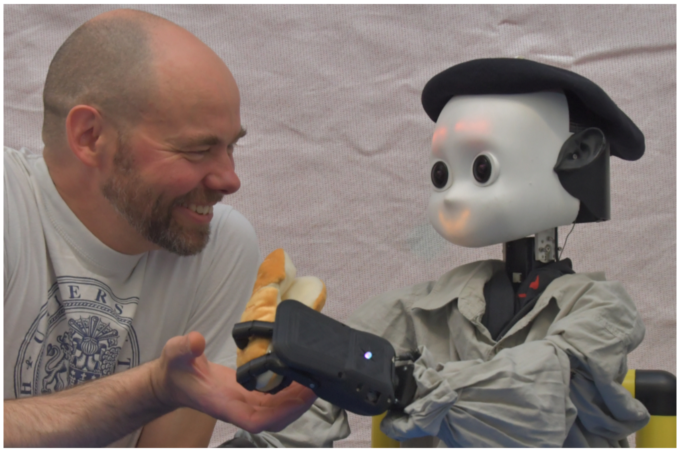


Fig. 6: The NICO robot [20] delivers a sandwich to a person. In this complex scenario robot should complete the given task without entering the human (personal) space and perform an object handover.

V. Conclusion

We presented an architecture for safe human-robot collaboration in a shared tabletop environment. Speech and gesture commands are used to instruct the robot on how to carry out its task without obstructing the human partner. Together with online tracking of the human hand, these instructions are translated into constraints on the motion planning of the robot. The combination of gesture and speech commands allows intuitive instruction of the robot, unlike a human collaborator would be instructed for a task. The motion constraints imposed by the human instructor can be seamlessly integrated into existing motion planning frameworks like ROS MoveIt!.

In future work, we will employ the approach on a humanoid robot in a collaborative tabletop scenario (see Figure 6). We are especially interested in exploring intuitive and reliable interface interaction strategies, which allow non-expert users to benefit from robot assistants in an intuitive, reliable and safe way. We will extend the approach with an extended speech and gesture instruction set, that allows the user more control over the behavior of the robot. Also, we will explore a more active role of the robot during collaboration; by integrating modules for action and plan recognition the robot will be able to assist the human user in a more proactive way.

VI. Acknowledgements

The authors would like to thank Erik Strahl for his support with the experimental setup and Hwei Geok Ng for her valuable advice and participation during prototype experiments.

References

[1] J. Peters and S. Schaal, "Learning to control in operational space," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 197–212, 2008.

[2] M. Vasic and A. Billard, "Safety issues in human-robot interactions," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 197–204.

[3] S. Schaal, "The new robotics–towards human-centered machines," *HFSP journal*, vol. 1, no. 2, pp. 115–126, 2007.

[4] S. Schaal and C. G. Atkeson, "Learning control in robotics," *IEEE Robotics & Automation Magazine*, vol. 17, no. 2, pp. 20–29, 2010.

[5] S. Lauria, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein, "Converting natural language route instructions into robot executable procedures," in *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*. IEEE, 2002, pp. 223–228.

[6] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein, "Mobile robot programming using natural language," *Robotics and Autonomous Systems*, vol. 38, no. 3, pp. 171–181, 2002.

[7] T. Nishizawa, K. Kishita, Y. Takano, Y. Fujita, *et al.*, "Proposed system of unlocking potentially hazardous function of robot based on verbal communication," in *System Integration (SII), 2011 IEEE/SICE International Symposium on*. IEEE, 2011, pp. 1208–1213.

[8] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Short text understanding through lexical-semantic analysis," in *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 2015, pp. 495–506.

[9] A. Abdulkader, A. Lakshmiratan, and J. Zhang, "Introducing DeepText : Facebook ' s text understanding engine," pp. 5–7, 2016.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[11] M. A. Zamani, S. Magg, C. Weber, and S. Wermter, "Deep reinforcement learning using symbolic representation for performing spoken language instructions." *In 2nd Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics (BAILAR) on Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on.*, 2017.

[12] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "Emorl: Continuous acoustic emotion classification using deep reinforcement learning," *accepted at the International Conference on Robotics and Automation (ICRA)*, 2018.

[13] M. Kerzel, H. Beik Mohammadi, M. A. Zamani, and S. Wermter, "Accelerating deep continuous reinforcement learning through task simplification," Rio de Janeiro, Brazil, 2018.

[14] S. Otte, A. Zwiener, and M. V. Butz, "Inherently constraint-aware control of many-joint robot arms with inverse recurrent models," in *Artificial Neural Networks and Machine Learning – ICANN 2017*, A. Lintas, S. Rovetta, P. F. Verschure, and A. E. Villa, Eds. Cham: Springer International Publishing, 2017, pp. 262–270.

[15] G. Yen and T. Hickey, "Reinforcement learning algorithms for robotic navigation in dynamic environments," in *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, vol. 2, 2002, pp. 1444–1449.

[16] F. Lindner, "A conceptual model of personal space for human-aware robot activity placement," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 5770–5775.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop*, 2013.

[18] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Nov. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078186

[19] P. Beeson and B. Ames, "Trac-ik: An open-source library for improved solving of generic inverse kinematics," in *Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference on*. IEEE, 2015, pp. 928–935.

[20] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "Niconeuro-inspired companion: A developmental humanoid robot platform for multimodal interaction," in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, 2017, pp. 113–120.